

Figures 7 through 9 provide flowcharts depicting logic which may be used to implement preferred embodiments of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention goes beyond the prior art, which provides QoS controls per client/server pairing, per application, or per Web object, and ties QoS to a specific transaction and all the related Web objects comprising that transaction. Moreover, the present invention also provides transaction-based QoS in an environment in which Web objects are sometimes served by application servers and sometimes served by distributed Web caches, surrogates, and proxies, hereinafter called "edge servers": the novel techniques which are disclosed enable the transaction-based QoS to be performed at varying points within a network path, providing an extremely powerful and flexible solution. The disclosed techniques also allow for using transaction-based QoS within a hierarchy of Web application servers and edge servers, and permit heterogeneity of QoS policy definitions in a network and heterogeneous QoS handling within a particular application – all without a dependency on identifying a client and server by their IP address and port number combinations and without requiring clients (or client-side proxies) to support cookies.

The term "server site" as used herein refers to the collection of server nodes that serve Web content associated with a given fully-qualified domain name. Fig. 1 provides a diagram of a representative server site 100, which may (for purposes of example) serve content for a domain name such as "www.ibm.com". This example server site 100 comprises a cluster 150 of application servers 140 (such as IBM WebSphere® application servers); several back-end